DOCUMENT RESUME

ED 395 991 TM 025 185

AUTHOR Sireci, Stephen G.
TITLE Using Cluster Analysis To Solve the Problem of Standard Setting.
PUB DATE 14 Aug 95
NOTE 28p.; Paper presented at the Annual Meeting of the American Psychological Association (New York, NY, August 1995).
PUB TYPE Reports - Evaluative/Feasibility (142) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS *Classification; *Cluster Analysis; Equated Scores; High School Equivalency Programs; High Schools; *High School Seniors; *Performance Based Assessment; Responses; Standards; Writing (Composition); *Writing Tests
IDENTIFIERS Dimensionality (Tests); Experts; *Standard Setting

ABSTRACT

This paper presents and evaluates a new procedure to classify examinees into meaningful categories based on their responses to (performance on) items comprising a test instrument. The proposed procedure uses the data analytic technique of cluster analysis to partition the examinee population into homogeneous groupings. This procedure was applied to a high-stakes writing skills test. External self-reported grades in English composition classes were used to validate the cluster solutions. Two separate samples of 511 and 390 high school seniors from national equating samples for the Test of General Educational Development were analyzed to evaluate the stability of the obtained solutions. A five-cluster solution best separated the examinees according to their writing skills proficiency and led to a passing standard that was stable across the two samples. The benefits of using cluster analysis to set standards on tests include reduced subjectivity in comparison to methods based on expert ratings of items, applicability to performance-based assessments, and absence of dimensionality requirements. The limitations of the procedure include the need for external criterion data and the lack of research in this area. (Contains 10 tables and 13 references.) (Author/SLD)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

TM

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
☑ This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.
• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY
STEPHEN G. SIRECI
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

ED 395 991

Using Cluster Analysis to Solve the Problem of Standard Setting

Stephen G. Sireci[1]

University of Massachusetts at Amherst

Paper presented at the Annual Meeting of the American Psychological Association as part of the symposium "Standard Setting: A Mixed Bag of Judgment, Psychometrics, and Policy," New York, August 14, 1995

---

[1]Special thanks to Lynn Shelley-Sireci and Andrew Wiley for their comments on an earlier draft of this paper.

TM025185

BEST COPY AVAILABLE

Abstract

This paper presents and evaluates a new procedure to classify examinees into meaningful categories based on their responses to (performance on) items comprising a test instrument. The proposed procedure uses the data analytic technique of cluster analysis to partition the examinee population into homogeneous groupings. This procedure was applied to a high-stakes writing skills test. External self-reported grades in English Composition classes were used to validate the cluster solutions. Two separate samples of examinees were analyzed to evaluate the stability of the obtained solutions. A five-cluster solution best separated the examinees according to their writing skills proficiency and led to a passing standard that was stable across the two samples. The benefits of using cluster analysis to set standards on tests include reduced subjectivity in comparison to methods based on expert ratings of items, applicability to performance-based assessments, and absence of dimensionality requirements. The limitations of the procedure include the need for external criterion data, and the lack of research in this area.

Introduction

When educational or psychological tests are used to make decisions or diagnostic classifications about people, standards or cutscores are needed to classify examinees into groups. Examples of group classifications based on educational tests are "basic," "proficient," "advanced," "passers," "non-passers," "masters," "non-masters," and "competent." In psychological testing, classifications typically refer to descriptions of the attribute being measured such as "aggressive," "assertive," or "extraverted." In either case, "cutpoints" are defined along the test score scale to determine categories into which examinees are placed.

The process of standard setting is an inexact science. The goal is to minimize classification errors so that examinees are categorized according to their "true" group membership. However, standard setting has been severely criticized and many test specialists argue that it should be abandoned whenever possible (Glass, 1978; Green, 1981; Shepard, 1979). The criticisms against standard setting stem from the fact that the process is largely subjective. The most popular methods used to set standards are those based on "expert" judgment (cf. Sireci & Biskin, 1992). Popular examples are item-based methods such as the Angoff procedure where groups of subject matter experts are asked to determine the proportion of minimally competent (borderline) examinees who would answer a test item correctly. Although expert judgment drives the process, the resultant standards often do not generalize from one group of experts to another, and vary widely depending on the particular standard setting method employed.

The purpose of this paper is to present and evaluate a new method of standard setting designed to reduce the degree of subjectivity involved in the process. This method uses the data analytic technique of cluster analysis to partition the examinee population into homogeneous groups based on test performance. The benefit of this procedure is that the partitioning scheme is empirically-based, rather than founded on subjective, albeit expert, judgment.

Cluster Analysis and Standard Setting

Standard setting as classification

Standards are set on tests so that examinees can be classified into one of two or more groups according to their standing on the attribute measured. Although test scores may be considered theoretically to be on a continuum, when standards are employed, it is implicitly assumed that the examinee population comprises two or more subgroups. It is assumed further that an examinee's performance on the test can be used to determine to which group

an examinee belongs. A common example is the use of tests in licensure and credentialing. Those who score at or above the passing score receive a license/credential and those who do not reach this standard, do not. Thus the population of examinees is, in effect, stratified into at least two "types" of people: those who are competent and are worthy of passing, and those who are below the defined level of competence and are not worthy of passing.

Cluster analysis represents a family of mathematical models designed to classify objects into categories based on empirical criteria. Because making judgements about examinees based on their test scores invokes classification rules, it seems sensible to apply cluster analysis to the standard setting problem.

Classification using cluster analysis

The purpose of cluster analysis is to partition a data set into homogeneous subsets, called "clusters." There are many different cluster analysis models, but they are generally classifiable into two families: clustering or hierarchical models, and partitioning or K-means models. Hierarchical cluster analyses merge objects comprising a data set in stepwise fashion. At the initial stage, all objects represent a separate cluster. The two most similar objects are merged at each subsequent stage until all observations comprise a single cluster. Partitioning models group observations into a given number of categories specified in advance by the investigator. The most popular partitioning strategy is the K-means procedure where objects are partitioned according to their standing with respect to K variables. The K-means for a given cluster represent the coordinates (centroid) of the cluster in K-dimensional space.

K-means cluster analysis is particularly applicable to the standard setting problem because the test developer can specify the number of groups to be identified. For example, for pass versus fail decisions, a two cluster solution can be specified in order to classify examinees into either the "pass" or "fail" clusters; however, higher-order cluster solutions may be required to identify several "passing" or "failing" clusters. Hierarchical models would be unwieldy when moderate to large numbers of examinees are tested, due to the difficulty in interpreting cluster solutions comprising hundreds of clusters.

Although using a K-means procedure to set standards on tests is intuitively appealing, there are several problems to be addressed. The most conspicuous problem is that the procedure will group examinees into the number of clusters specified by the investigator regardless of whether true clusters exist in the examinee population. Therefore, in using cluster analysis to set standards on tests, examinee clusters must be validated against external criteria. For this reason, using cluster analysis to

"solve" the standard setting problem does not remove subjectivity from the process. Judgments must be used to determine adequate validity criteria and to establish rules for determining cluster membership (i.e., define the test score(s) that signify membership in a cluster). However, using expert judgment to evaluate cluster solutions is far less subjective than gathering expert judgments on a multitude of item ratings. Furthermore, the cluster solution provides information regarding the "true" composition of the examinee population in terms of proficiency-based groupings. A second problem in using cluster analysis to set standards is demonstrating the stability of the cluster solution over different samples of examinees.

This paper explores and evaluates the use of a K-means clustering procedure to set the passing standard on a high-stakes writing skills assessment used as part of a battery to credentialize a high school diploma. In evaluating the utility of this procedure, the issues of passing standard validity and stability (reliability) are considered.

Method

Instrument

The examination investigated in this study is the Writing Skills Test (WST) of the Tests of General Educational Development (GED Tests). The GED Tests are a battery of five tests (Writing Skills, Social Studies, Science, Interpreting Literature and the Arts, and Mathematics) used to grant a high school-level diploma to adults who did not complete high school. By passing all five tests in the GED battery, adults can further their educational, occupational, or other personal goals. The current procedure for establishing passing scores on the GED Tests is a norm-referenced procedure. However, the norm group used to set the passing standards that adult GED candidates must meet is a nationally-representative sample of graduating high school seniors. Thus, GED passing scores are norm-referenced using a criterion reference group. In order for adults to pass the GED Tests, they must demonstrate a level of performance on the test battery that is equal to or greater than the 25th percentile of U.S. graduating high school seniors (GEDTS, 1993).

The WST comprises 50 multiple-choice (MC) questions and a single essay. The content areas measured by the MC section are sentence structure (35%) usage (35%), and mechanics (30%). The essay portion of the WST measures a GED candidate's ability to compose a well-written response to an essay topic. Examinees are presented with an expository topic (e.g., state a view or present an opinion) and asked to write a response in essay form. GED candidates have two hours to complete the WST. They are advised to use 75 minutes for the MC questions and 45 minutes for the essay. The essay is scored holistically on a six-point scale by

two highly-trained readers. If the separate scores assigned by the two readers differ by more than one point, the essay is scored by a chief reader. The total essay score is the sum of the readers' scores when graded by two readers, and twice the average of the three scores when graded by three readers. Thus, the essay scores range from two to twelve. Estimates of interrater reliability of the total WST essay score are high (around .91; Wiley & Sireci, 1994).

The essay and MC sections of the WST are weighted .36 and .64, respectively, and scaled to form a composite score. The composite score scale ranges from a minimum of 20 to a maximum of 80, with a standard deviation of 10 scale score units. Previous research has found the reliability of WST composite scores to be around .87 (Patience & Swartz, 1987; Lukhele & Sireci, 1995; Wiley & Sireci, 1994).

Subjects

The data used in this study come from two national equating studies of new GED test forms. The analyses reported in this paper focus on 511 high school seniors who took one form of the Writing Skills Test (WST) in 1994, and 390 high school seniors who took this same form in 1993. In addition to their test scores, self-reported grades in English Composition courses were obtained.

Procedure

Cluster analysis

Four separate test scores were used to separately classify the 1994 and 1993 samples into clusters. In addition to their standardized essay scores, three separate content area scores were computed for all students. These additional content area scores were computed by summing examinees' dichotomously-scored responses to MC items comprising the sentence structure, usage, and mechanics content areas. Thus the four variables used to cluster examinees were: essay scores, sentence structure scores, usage scores, and mechanics scores. All scores were standardized to a z-score scale (mean=0, s.d.=1) prior to clustering.

A K-means clustering procedure was used to classify examinees into clusters according to their similarity with respect to the four WST subscores (thus K=4). Two-through eight-cluster solutions were obtained (i.e., specified *a priori*) for each sample. Higher cluster solutions were not attempted because the eight-cluster solution exhibited some clusters comprising fewer than 30 examinees. The 1994 sample was used to generate the initial cluster solutions; the 1993 sample was used to evaluate the stability of the 1994 results.

The K-means clustering algorithm. K-means cluster analysis is an iterative process. All analyses reported in this paper were conducted using SPSS for Windows, version 6.0 (SPSS, 1993). The K-means algorithm incorporated into SPSS is Anderberg's (1973) nearest centroid sorting method. This method calculates the Euclidean distance of each object (examinee) to the center (centroid) of every cluster, and assigns the examinee to the cluster which it is "closest" (i.e., to which its Euclidean distance from the centroid is smallest). Anderberg's procedure minimizes the variance within a cluster and maximizes the variance between clusters. For a "C"-cluster solution, where C is the number of clusters requested, the initial stage uses the values (scores) of the C examinees who differ most from one another as the initial estimates of the cluster centers (elements of the centroid vector). The Euclidean distance of all examinees to these centroids is computed and examinees are assigned to their "closest" cluster. The cluster centroids are then re-calculated (based on examinees comprising the cluster) and new distances are calculated for all examinees. New cluster assignments are then made based on the new distances. This process repeats until no changes in cluster membership occur, or until some other convergence criterion is reached. The convergence criterion used here was a maximum distance change of less than .02 by any cluster center between two sequential iterations. That is, if the change in distances of all cluster centers from one iteration to the next was less than 2%, the solution was considered to be converged.

Validating the cluster solutions

Because the K-means procedure minimizes within-cluster variance and maximizes between cluster variance, the clusters that emerge typically differ substantially on the variables used to create the clusters. A cursory check that the clusters do differ significantly can be accomplished by comparing the cluster means on these variables individually using t-tests (for the two-cluster case) or ANOVAs (for higher cluster solutions); or by comparing the cluster centroids across all variables simultaneously using MANOVA procedures. However, given the "min-max" objective of cluster analysis, such analyses are not particularly meaningful, and do not validate that the obtained clusters represent "real" groups (Aldenderfer & Blashfield, 1984). Therefore external criteria (i.e., variables that were not used to form the clusters) are needed to validate that the groups differ with respect to the construct measured, rather than due to statistical artifact.

In this study, self-reported grades in English Composition courses were used as the external criteria to evaluate the cluster solutions. These grades were taken as an alternative measure of the students' writing skills proficiency. Therefore, for the clusters within a solution to be valid (i.e., represent

groups that truly differ with respect to writing proficiency) they should also differ with respect to English Composition grades. ANOVAs and t-tests were conducted to compare differences in standardized English Composition grades across clusters comprising a given solution (e.g., do the four groups obtained via a four-cluster solution differ with respect to English Composition grades?).

Selecting the most appropriate cluster solution

English composition grades were also used to help identify the clustering solution that best dichotomized the examinee sample into "passers" and "failers." Point-biserial, t-test, and cross-tabulation analyses were carried out to evaluate competing cluster solutions. Initially, cross-tabulations were computed for each cluster solution stratifying by cluster membership and English Composition grades. The different cluster solutions were compared by determining the percentage of "A" students who were placed in "failing" clusters and the percentage of "D or F" students who were placed in passing clusters. Failing "A" students and passing "D or F" students were considered to be classification errors, and so the purpose of this comparison was to find the cluster solution that best minimized these errors. Following this analysis, the clusters within a solution were individually classified as "passing" clusters or "failing" clusters. New cross-tabulations were generated exploring the relationship between "A" and "D/F" students and pass/fail status.

Students were assigned two separate pass (1) or fail (0) classifications: the first based on their cluster membership, the second was based on the median WST score attained by the "borderline" cluster. Point-biserial correlations were obtained between pass/fail status within a specific cluster solution and English Composition grades. The purpose of this comparison was to identify the cluster solution that generated the largest correlation between pass/fail status and English Composition grades. Finally, separate t-tests were conducted for each cluster solution using pass/fail status as the grouping variable, and standardized English Composition grades as the dependent variable. These analyses sought to identify cluster solutions that provided pass/fail classifications that best discriminated between the groups with respect to grades in English Composition.

Evaluating the stability of the cluster solutions

To determine whether the obtained cluster solutions were stable across different samples of high school seniors, the cluster analyses and follow-up comparisons with English Composition grades were repeated using the 1993 sample of high school seniors. Similarity of the cluster profiles, passing standards, and passing percentages resulting from the 1993 and 1994 samples were evaluated.

Results

Cluster Analyses: 1994 Sample

Two- through eight-cluster solutions were obtained for the 511 students who took the WST in 1994. The eight-cluster solution exhibited some groups comprising fewer than 30 examinees, and so higher-cluster solutions were not attempted. The proportion of students in each cluster, and the centroids for each group (i.e., cluster means on the four standardized variables used to cluster students) are presented in Table 1. Also listed in Table 1 are the mean and median WST composite scores for all students falling within a cluster.

The two-cluster solution separated the students into a high-scoring group and a low scoring group. In terms of mean composite score, these two groups were over one-and-a-half standard deviations apart. The low-scoring group had negative mean scores across the four standardized variables (essay, mechanics, sentence structure, and usage scores), while the high-scoring group had positive scores. A high-scoring and low-scoring group were also observed in the three-cluster solution, in addition to a group who scored relatively poorly on the essay, in comparison to the MC subscores.

The four-cluster solution revealed two groups that scored relatively low across all four variables. These two groups differed by about .5 to 1.3 standard deviation units across the four variables. The other two groups represented a high scoring group, and a group that performed below average on the essay, but above average on the MC sections.

The five-cluster solution revealed three low-scoring groups: one group exhibited extreme low scores across all four variables; one group slightly below average on the essay, but extremely below average on the MC sections; and one group moderately below average across all four variables. The other two groups represented a group who scored well above average on all four variables, and a group who scored below average on the essay, but above average on the three MC sections.

The six-cluster solution was similar to the five-cluster solution, with the sixth cluster representing an additional high-scoring group. That is, three groups were below average on all four variables, two groups were above average across all variables, and one group exhibited below average essay scores and above average MC scores.

Table 1

Summary of Cluster Analysis Results for 1994 Sample

Solution	Clustr	%	Mean Compos.	Median Compos.	Essay	Mech.	Sent. Struc.	Usage
2-cluster	2	34	38.58	40	-.70	-1.07	-1.12	-1.11
	1	66	54.05	53	.36	.56	.58	.58
3-cluster	1	25	37.04	38	-.72	-1.36	-1.36	-1.41
	3	41	48.00	48	-.42	.25	.25	.28
	2	33	58.66	57	1.07	.73	.73	.73
4-cluster	1	13	33.11	34	-.98	-1.90	-1.79	-1.88
	2	24	42.65	43	-.36	-.54	-.69	-.60
	4	32	49.77	49	-.43	.47	.54	.48
	3	31	58.91	58	1.14	.72	.72	.74
5-cluster	1	8	32.38	34	-1.74	-1.62	-1.49	-1.50
	3	13	38.19	39	-.18	-1.40	-1.56	-1.63
	4	26	44.77	45	-.27	-.27	-.25	-.18
	2	24	51.52	51	-.40	.66	.70	.63
	5	29	59.44	58.5	1.17	.78	.75	.78
6-cluster	3	11	33.79	35	-.66	-1.90	-1.84	-2.00
	1	5	34.73	36	-1.81	-1.21	-1.21	-1.06
	4	19	43.53	43	-.15	-.54	-.69	-.57
	5	25	47.33	47	-.56	.29	.35	.31
	6	32	56.33	55	.56	.71	.73	.72
	2	8	64.90	64	1.98	.90	.90	.92
7-cluster	2	10	32.66	34	-.81	-2.02	-1.93	-2.04
	4	6	35.76	36	-1.79	-1.02	-1.04	-.83
	3	9	41.65	41	-.18	-.78	-1.01	-1.15
	7	21	45.40	45	-.15	-.25	-.23	-.11
	1	22	50.50	50	-.50	.58	.66	.58
	6	25	56.97	55	.77	.74	.71	.71
	5	7	65.74	64	2.02	.95	.95	.96
8-cluster	6	5	30.21	31	-1.84	-2.00	-1.69	-1.77
	2	9	36.27	37	-.26	-1.66	-1.79	-1.92
	1	5	39.22	39	-1.74	-.22	-.50	-.26
	5	18	42.75	43	-.42	-.49	-.56	-.55
	3	7	49.53	50	.89	-.30	-.23	-.08
	8	27	50.08	50	-.40	.49	.55	.53
	7	23	57.46	56	.77	.78	.76	.75
	4	7	66.00	64	2.01	.96	.98	.99

Notes: "%" indicates the percentage of students in each cluster. "Compos." indicates the WST composite score for examinees comprising the cluster on the 20-80 GED score scale. The other values indicate standardized (z-score) cluster means for essay, mechanics, sentence structure, and usage scores.

The seven-cluster solution revealed four groups that were below average across all four variables, two groups that were above average across all four variables, and a group that performed below-average on the essay, but above average on the MC sections. The eight-cluster solution identified similar groups (i.e., four below-average, two above-average, and one "poor essay"), and included a group that performed above average on the essay, but below average on the MC sections.

As expected, ANOVAs using scores on the essay, mechanics, sentence structure, and usage subscores as the independent variable, and cluster membership as the dependent variable exhibited statistical significance across all seven cluster solutions. Because these results are not surprising, they are not reported in detail here.

Evaluating the cluster solutions

Inspection of the mean or median composite score and the subscores for each cluster provides intuitive information about the relative strengths and weaknesses of students comprising the various clusters. It also provides estimates of the proportion of students who would pass the WST, if a given cluster was chosen as the "borderline group." Using only this information, a clustering solution could be chosen that represented a passing score targeted to a particular passing percentage. However, such norm-referenced information does not tell us whether the cluster solutions are meaningful with respect to "true" differences in writing proficiency. Furthermore, these data do not provide us with any concrete information regarding which cluster solution is most appropriate. To evaluate the validity of the clustering solutions, and to identify the solution that would best separate those who are proficient from those who are not, the students' self-reported English Composition grades were used. Grades were available for 461 of these 511 seniors (90%).

Preliminary analyses were conducted to see if the individual groups within a cluster solution (i.e., the different clusters) differed significantly with respect to English Composition grades. A t-test was used for the two-cluster group and one-way ANOVAs were conducted for the higher-order cluster solutions[2]. The t- and F-statistics were statistically significant (at $p \leq .01$) for all comparisons, indicating that the separate clusters

[2]It can be argued that because the English Composition grades represent an ordinal scale, nonparametric analyses using unstandardized grades would be more appropriate than parametreic analyses using standardized grades. Kruskal-Wallis ANOVAs and Mann-Whitney U-tests were also conducted using unstandardized grades to determine whether choice of statistical procedure would affect the results. No differences in statistical significance between the parametric and non-parametric analyses were observed, and so only the parametric analyses are reported.

within a solution differed with respect to English Composition grades. Post-hoc analyses illustrated which groups were statistically significantly different. These results are not presented here for reasons of brevity. The statistically significant t- and F-statistics verify that the cluster solutions did group examinees meaningfully in terms of their writing skills proficiency.

To determine which of the seven cluster solutions best discriminated "passers" from "non-passers," cross-tabulations, point-biserial correlations, and t-tests using English Composition grades as the dependent variable, were conducted. For the biserial and t-test analyses, the clusters in each solution were classified as either "passing clusters" or "failing clusters" according to their mean values on the standardized WST subscores. For example, in the five cluster solution, clusters 1, 3, and 4 were considered to comprise failing students, with cluster 4 representing the "borderline" group (see Table 1). The median WST composite score for examinees comprising the borderline group was taken as the passing standard suggested by the particular cluster solution. Table 2 lists the passing percentages for each solution based on this passing standard. The specific clusters classified as passing or failing, and the borderline cluster, are also identified in Table 2. Three of the seven cluster solutions resulted in the same passing standard (43). The passing percentages reported in Table 2 range from 66.5% for the five-cluster solution, to 87.5% for the three-cluster solution. It is interesting to note that the median (and mode!) passing percentage is 74.6%, which is close to the "25th percentile" standard to which the GED test battery is targeted.

Table 2

Passing Percentages, "Passing" and "Failing" Clusters, and Borderline Cluster: 1994 Sample (n=511)

Cluster Solution	Passing Clusters	Failing Clusters	Borderline Cluster	Passing Standard	Percentage Passing
3-Cluster	1,2	3	3	38	87.5
2-Cluster	1	2	2	40	83.6
7-Cluster	1,5,6,7	2,3,4	3	41	81.6
4-Cluster	1,4	2,3	3	43	74.6
8-Cluster	3,4,7,8	1,2,5,6	5	43	74.6
6-Cluster	1,4,5	2,3,6	4	43	74.6
5-Cluster	2,5	1,3,4	4	45	66.5

The results of the cross-tabulation analysis are summarized in Table 3. These results help evaluate the "misclassification rates" for the different cluster solutions. To make this evaluation, we define a "type 1" error as "passing a student who really should fail" and a "type 2" error as "failing a student who really should pass," and assume that all "A students" should pass and all "D or F" students should fail. Using these definitions, the five-cluster solution minimized type 1 errors (27.78% of "D/F" students pass) and the seven-cluster solution minimized type 2 errors (5.45% of the "A" students fail). The seven-cluster solution had the lowest overall misclassification rate (10.16%); improving classification for an additional eight students in comparison to the five-cluster solution.

Table 3

Percentage of "A Students" Failing and "D or F Students" Passing **Using Cluster Membership** As Passing Standard
(n=461)

Solution	"A Students" Failing %		"D/F Students" Passing %		Total Missclas. %	
	n	%	n	%	n	%
7-cluster	6	5.45	7	38.89	13	10.16
3-cluster	7	6.36	7	38.89	14	10.94
2-cluster	9	8.18	7	38.89	16	12.50
4,6-cluster	11	10.00	6	33.33	17	13.28
8-cluster	10	9.09	9	50.00	19	14.84
5-cluster	16	14.54	5	27.78	21	16.41

Note: Of the 460 students who reported English Composition grades, 110 were "A students," 1 was a "F" student, and 17 were "D" students. The percentages reported are based on total n's of 110, 18, and 128 for the "A," "D/F," and "total" groups, respectively. The percentages for the four- and six-cluster solutions were identical.

Table 4 presents the cross-tabulation results using the median composite score of the borderline cluster to classify examinees as pass/fail. After employing this passing standard for each cluster solution, the type 1 error rates (failing "A" students) decrease for all solutions and the type 2 error rates (passing "D/F" students) increase. The two- and three-cluster solutions, rather than the seven-cluster solution, exhibited the lowest type 1 error rate (2.72%). The five-cluster solution retained the lowest type 2 error rate (50%). The overall misclassification rates were 13.28% for all solutions except the five-cluster

solution, which was slightly higher due to one additional misclassification. The overall misclassification rates increased for the two- , three-, and seven-cluster solution and decreased for the five-cluster solution, in comparison to the cross-tabulation analysis based on pass/fail cluster membership.

Table 4

Percentage of "A Students" Failing and "D or F Students" Passing **Using Borderline Cluster's Median** as Passing Standard 1994 Sample (n=461)

Solution	"A Students" Failing		"D/F Students" Passing		Total Missclas.	
	n	%	n	%	n	%
7-cluster	4	3.64	13	72.22	17	13.28
2,3-cluster	3	2.72	14	77.78	17	13.28
4,6,8-cluster	7	6.36	10	55.56	17	13.28
5-cluster	9	8.18	9	50.00	18	14.06

Note: Of the 460 students who reported English Composition grades, 110 were "A students," 1 was a "F" student, and 17 were "D" students. The percentages reported are based on total n's of 110, 18, and 128 for the "A," "D/F," and "total" groups, respectively. The percentages for the two- and three-cluster solutions were identical, as were those for the four-, six-, and eight-cluster solutions.

Mean differences in standardized English Composition grades between "passers" and "failers" in each of the 7 cluster solutions are presented in Table 5, along with the group means, standard deviations, and t-statistics. The mean differences ranged from .70 to .83 standard deviation units; all were statistically significant. This finding indicates that the pass/fail classifications based on the median composite score of the borderline cluster significantly distinguished between those who received relatively higher English Composition grades and those who received relatively lower English Composition grades. In comparing the solutions, the five-cluster solution exhibited the largest mean difference in grades between the pass and fail groups, and had the largest t-value. Thus when using all grades rather than just grades of "A" and "F/D," the five-cluster solution best separated students according to their English Composition grades.

Two separate point-biserial correlations are presented in Table 6. These separate correlations reflect the relationship between English Composition grades and pass/fail classifications using the two different methods for classifying examinees as "pass" or "fail." The first method used actual cluster membership; that

is, students who were grouped together in "failing clusters" (see Table 2) were classified as"fail," while the others were classified as "pass." The second method used the median WST composite score for the borderline cluster. Students who scored at or above this median were classified as "pass," those who fell below this median were classified as "fail." The correlations based on cluster membership are useful for evaluating the relationship between the cluster solutions and grades, while the correlations based on median composite score are informative for evaluating the passing standards offered by the different clustering solutions.

Table 5

Comparing Mean Standardized English Composition Grades Passers Versus Failers Across Seven Cluster Solutions 1994 Sample (n=461)

Cluster Solution	Passing Seniors			Failing Seniors				
	n	X	s.d.	n	X	s.d.	X Dif.	t
3-cluster	417	.07	.97	44	-.70	1.00	.77	4.99*
2-cluster	408	.09	.98	53	-.61	.92	.70	5.29*
7-cluster	374	.11	.97	87	-.59	.93	.70	6.75*
4,6,8-cls.	358	.17	.95	103	-.58	.95	.75	7.01*
5-cluster	318	.25	.94	143	-.58	.90	.83	8.65*

*$p \leq .01$

The point-biserials reflecting pass/fail classifications based on cluster membership exhibited larger correlations with English Composition grades than did those reflecting classifications based on reaching the median score of the borderline cluster. These correlations ranged from .334 (3-cluster solution) to .422 (five-cluster solution), and accounted for between 11 and 18 percent of the variance of the English Composition grades. Testing for significant differences between the correlations for each solution (using Fisher's Z) indicated that only the three-cluster solution was statistically significantly different from the five-cluster and two-cluster solutions.

The point-biserials reflecting pass/fail classifications based on reaching the median composite score for the borderline clusters ranged from .227 to .374, and accounted for between five and fourteen percent of the variance of the English Composition

grades. Using Fisher's Z, the correlation exhibited by the five-cluster solution was statistically larger than those exhibited by the two, three-, and seven-cluster solutions. It was not statistically significantly larger than the correlation resulting from the four-, six, and eight-cluster solutions (passing WST score of 43). The squared point-biserials reported for the borderline cluster classifications also serve as estimates of the effect size for the t-tests reported in Table 5.

Table 6

Correlations Between Pass/Fail Status and Eng. Composition Grades
1994 Sample (n=461)

Solution	Pass/Fail Classification Based on: Cluster Membership		Borderline Median	
	r_{pb}	r^2_{pb}	r_{pb}	r^2_{pb}
3-cluster	.334	.112	.227	.052
7-cluster	.350	.123	.250	.063
4-cluster	.378	.143	.311	.097
6-cluster	.360	.130	.311	.097
8-cluster	.375	.141	.311	.097
2-cluster	.388	.151	.240	.056
5-cluster	.422	.178	.374	.140

To summarize the initial clustering results, the analyses comparing pass/fail classifications and English Composition grades indicated that the groupings of students provided by the cluster solutions were significantly related to their English Composition grades. In evaluating the passing standards offered by the different cluster solutions, the five-cluster solution stands out because it exhibited the strongest relationship with English Composition grades and represented the most rigorous passing standard (WST score of 45). For these reasons, it seems to be the solution that best reflects the "true" structure of the examinee population. A case can also be made to use a score of 43 as the passing standard because this score was advanced by three of the seven clustering solutions, and its corresponding pass/fail classifications were related to English Composition grades in similar fashion to the five-cluster solution. Therefore, the initial clustering results support a passing score in the 43-to-45 scale score range.

Stability of the Cluster Solution: Clustering the 1993 Sample

The analyses reported above illustrate that the clustering solutions were valid for grouping students with respect to writing skills proficiency. This section explores the stability of the obtained cluster solution. Two pertinent questions addressed are:

1) If a different sample of examinees were used, would the same clusters emerge?

2) If a different sample of examinees were used, would we arrive at the same passing standard?

To answer these questions, the clustering procedure was applied to a sample of 390 graduating high school seniors who took the same form of the WST in 1993. Self-reported data on English Composition grades were available for 368 (94%) of these seniors.

Cluster analysis results

The results of the two- through eight-cluster solutions are presented in Table 7. The results for the 1994 sample are also presented in Table 7 to compare the types of groupings that emerged using the different data sets. At first glance, it appears that the same "types" of clusters emerged using both data sets. For example, the four-cluster solution for both samples comprises two low-scoring clusters, one high-scoring cluster, and a cluster that performed below average on the essay, but above average on the MC sections. Upon closer inspection, it can be seen that the 1993 clusters tended to differ more with respect to the MC subscores than did the 1994 sample. The MC subscores tended to operate as a set using the 1994 data. Using the 1993 data, the MC subscores appear to have more influence upon the clustering solution individually. Although this difference is notable, the median composite scores for the groups comprising the cluster solutions are fairly similar across the 1993 and 1994 samples. In 29 out of the 35 cases (83%), these differences are within three standard score points, and almost half (46%) were within a single point. The profiles of the groups within a clustering solution are very similar, especially for the groups that would be labeled "borderline" clusters.

To determine the impact of the different samples on the standards that would be set on the WST using each group, the passing standards resulting from the different samples were compared. This comparison evaluated the "stability" of the resulting passing score, as opposed to the stability of the cluster solutions. These "passing standards" are presented in Table 8. Also presented in this table are the percentage of students who would meet these standards and the difference between the passing percentages. The passing percentages reported in Table 8 were

BEST COPY AVAILABLE

Table 7: Comparison of Cluster Solutions: 1994 & 1993

	1994 (n=511, X=48.76)						**1993** (n=390, X=49.32)					
Solution	%	Cmp.Mn.	Essay	Mech	SntStr	Use	%	Cmp.Mn.	Essay	Mech	SntStr	Use
2-Cluster	34	40	-.70	-1.07	-1.12	-1.11	24	37	-.90	-1.39	-1.47	-1.47
	66	53	.36	.56	.58	.58	76	53	.29	.44	.47	.47
3-Cluster	25	38	-.72	-1.36	-1.36	-1.41	19	36	-.92	-1.61	-1.67	-1.69
	41	48	-.42	.25	.25	.28	42	47	-.43	.13	.14	.16
	33	57	1.07	.73	.73	.73	39	57	.93	.66	.69	.67
4-Cluster	13	34	-.98	-1.90	-1.79	-1.88	12	32	-1.08	-2.06	-1.90	-2.08
	24	43	-.36	-.54	-.69	-.60	15	41	-.58	-.74	-1.01	-.86
	32	49	-.43	.47	.54	.48	39	49	-.34	.27	.28	.31
	31	58	1.14	.72	.72	.74	35	59	1.00	.69	.74	.71
5-Cluster	8	34	-1.74	-1.62	-1.49	-1.50	15	34	-.97	-1.78	-1.84	-1.91
	13	39	-.18	-1.40	-1.56	-1.63	9	42.5	-1.22	.12	-.35	-.22
	26	45	-.27	-.27	-.25	-.18	19	45	-.11	-.53	-.32	-.29
	24	51	-.40	.66	.70	.63	45	54	.18	.60	.63	.62
	29	58.5	1.17	.78	.75	.78	11	63	1.78	.78	.81	.79
6-Cluster	11	35	-.66	-1.90	-1.84	-2.00	6	30	-1.85	-1.71	-1.94	-1.95
	5	36	-1.81	-1.21	-1.21	-1.06	8	36	-.35	-2.00	-1.80	-1.98
	19	43	-.15	-.54	-.69	-.57	11	41	-.81	-.67	-.91	-.74
	25	47	-.56	.29	.35	.31	23	47	.03	-.10	-.05	-.01
	32	55	.56	.71	.73	.72	22	52	-.49	.65	.65	.63
	8	64	1.98	.90	.90	.92	30	59	1.10	.72	.76	.73
7 Cluster	10	34	-.81	-2.02	-1.93	-2.04	5	30	-1.92	-1.77	-1.95	-2.01
	6	36	-1.79	-1.02	-1.04	-.83	7	36	-.34	-2.09	-1.84	-2.03
	9	41	-.18	-.78	-1.01	-1.15	11	40	-.78	-.77	-1.04	-.87
	21	45	-.15	-.25	-.23	-.11	4	42	-1.67	.39	.29	.17
	22	50	-.50	.58	.66	.58	24	47	-.08	-.14	-.08	.04
	25	55	.77	.74	.71	.71	38	55	.24	.70	.71	.65
	7	64	2.02	.95	.95	.96	11	63	1.78	.78	.81	.79
8 Cluster	5	31	-1.84	-2.00	-1.69	-1.77	4	29	-1.98	-2.14	-1.99	-2.13
	9	37	-.26	-1.66	-1.79	-1.92	4	32	-.44	-2.40	-2.07	-2.33
	5	39	-1.74	-.22	-.50	-.26	11	39	-.77	-1.09	-1.39	-1.31
	18	43	-.42	-.49	-.56	-.55	4	42	-1.72	.26	.13	-.01
	7	50	.89	-.30	-.23	-.08	19	45	-.46	-.20	-.19	.05
	27	50	-.40	.49	.55	.53	12	50.5	.60	-.02	.03	-.09
	23	56	.77	.78	.76	.75	35	55	.19	.72	.75	.70
	7	64	2.01	.96	.98	.99	11	63	1.79	.80	.82	.80

BEST COPY AVAILABLE

calculated by combining the 1993 and 1994 samples, so that the passing percentages would not be sample-specific.

In terms of the passing standard (i.e., borderline cluster median), only the five-cluster solution resulted in the same passing score for both samples. This finding is noteworthy because the five-cluster solution was deemed to be the "correct" cluster solution using the 1994 data. The next "most stable" solutions were the seven- and eight-cluster solutions, which differed by one point across 1993 and 1994. The difference in passing percentage between 1993 and 1994 was lower for the seven-cluster solution (1.4%), than it was for the eight-cluster solution (3.1%). This finding is due to the fact that more examinees fell within the 42-to-43 scale score range than within the 40-to-41 scale score range. For the other solutions, the passing standard was lowered by one to three WST scale score points. The four- and six-cluster solutions again provided the identical passing standard using the 1993 data (WST score of 41), but this score was two points lower than the standard they provided using the 1994 data (43, which was also suggested by the eight-cluster solution).

Table 8

1993 and 1994 Passing Standards
and Combined-Sample Passing Percentages

Cluster Solution	1994 Standard	1993/1994 % Passing	1993 Standard	1993/1994 % Passing	% Diffr.
2-Cluster	40	83.2	37	89.0	5.8
3-Cluster	38	87.5	36	91.1	3.6
4-Cluster	43	75.7	41	81.8	6.1
5-Cluster	45	68.5	45	68.5	0.0
6-Cluster	43	75.7	41	81.8	6.1
7-Cluster	41	81.8	40	83.2	1.4
8-Cluster	43	75.7	42	78.8	3.1

Note: "1993/1994 % Passing" indicates the percentage of students in the combined 1993/1994 data set that would reach the standard listed.

Analyses of the 1993 cluster solution with respect to English Composition grades also supported the stability of the five-cluster solution over the two samples. Table 9 presents the point-biserial correlations between pass/fail classifications based on the borderline cluster's median composite score and English Composition grades. The observed correlations were slightly higher than those obtained from the 1994 sample. The five-cluster solution once again exhibited the largest correlation. Testing for significant differences between these correlations using Fisher's z-test indicated that the correlation based on the five-cluster solution was statistically significantly larger than those based on the two- and three-cluster solutions. As can be inferred from the magnitude of these correlations, the five-cluster solution also exhibited the largest t-value when the mean English Composition grades were compared across the pass/fail groups.

Table 9

English Composition Pass/Fail Point-Biserial Correlations
1993 Sample (n=367)

Cluster Solution	r_{pb}	r^2_{pb}
3-Cluster	.235	.056
2-Cluster	.273	.075
6-Cluster	.321	.103
7-Cluster	.321	.103
8-Cluster	.327	.107
4-Cluster	.329	.108
5-Cluster	.388	.151

Cross-tabulation analyses of the pass/fail classifications based on each cluster solution and "A" versus "D/F" students are summarized in Table 10. The overall misclassification rates for all cluster solutions were lower than those observed using the 1994 data. This finding (and the larger correlations observed for the 1993 data) may stem from the fact that self-reported grades were available for a larger percentage of the 1993 sample (94%) than were available for the 1994 sample (90%). However, in both cases, the number of "D/F" students was low. The ranking of the classification rates for the five-cluster solution was identical to its ranking using the 1994 data: it "passed" the fewest "D/F" students, but failed the most "A" students. Unlike the 1994 analysis, the two- and three- cluster solutions failed the fewest "A" students. The six- and seven-cluster solutions had the lowest overall misclassification rates. These two solutions improved classification for an additional three students in comparison to the five-cluster solution.

Table 10

Percentage of "A Students" Failing and "D or F Students" Passing Using Borderline Cluster Mean as Passing Standard 1993 Sample (n=367)

Solution	"A Students" Failing		"D/F Students" Passing		Total Missclas.	
	n	%	n	%	n	%
6,7-cluster	3	2.75	8	53.33	11	8.87
2-cluster	3	2.75	9	60.00	12	9.67
3-cluster	2	1.83	10	66.67	12	9.67
4-cluster	4	3.67	8	53.33	12	9.67
8-cluster	5	4.59	8	53.33	13	10.48
5-cluster	8	7.33	6	40.00	14	11.29

Note: Of the 367 students who reported English Composition grades, 109 were "A students," 2 were "F" students, and 13 were "D" students. The percentages reported are based on total n's of 109, 15, and 124 for the "A," "D/F," and "total" groups, respectively. The passing percentages for the two- and three-cluster solutions, and for the six- and seven-cluster solutions were identical.

Discussion

The results of this study illustrate that it is possible to set standards on tests using cluster analysis. As with any standard setting method, the proposed procedure is not completely objective (Zieky, 1994). Judgment is required to select the most appropriate cluster solution, to determine how to use the clustering results to set the passing standard, and in deciding which types of misclassification errors are most damaging. However, because these types of judgments are made in concert with the empirical findings of the procedure, the overall level of subjectivity involved using cluster analysis to set standards is far below that required when convening panels of experts to rate individual test items or tasks.

To facilitate interpretation of the clustering results, subject matter experts (SMEs) should be used to evaluate the solution. The cluster analysis procedure does not obviate the need for "non-psychometric" experts, although the task for the SMEs in helping to interpret the clustering solutions is likely to be easier, and more interesting, than traditional item-rating tasks. Thus, the same type of experts that are used in item-rating standard setting studies should be employed. For example, SMEs could be used to determine whether different sections of a test

should be given more emphasis when deciding on the borderline cluster. With respect to the WST, a panel of writing skills SMEs may decide that the essay portion of the test should be given more weight, and select a cluster that exhibits higher performance on the essay to represent the borderline group.

Implications of the study for the WST

The results of this study do not purport to sanction a particular passing score for the WST. Rather, the WST was used to exemplify the proposed standard setting procedure. Interpreting these results in concert with a carefully-selected panel of SMEs, and more fully evaluating the stability of the results over different samples, would be required to recommend a passing standard for the WST. Identification of better external criteria, in particular actual (rather than self-reported) English Composition grades, would enhance the validity of the results[3]. Nonetheless, it is interesting to note that the most plausible passing standard identified in this study (i.e., standard score of 45) represents the 32nd percentile of the GED norm group, which is just above the 25-30 average (across the five-test battery) percentile rank range recommended as the minimum passing standard by the GED Testing Service (GEDTS, 1994).

Strengths and Weaknesses of the Clustering Procedure

The results of this study illustrate the need for one or more strong external criteria to verify that the clustering results are meaningful. Without such criteria, there is no way of knowing for sure that the clustering solutions represent "real" examinee groupings. This study employed a single criterion to evaluate the cluster solutions. Using multiple criteria would probably enhance the validity of the results. Although the use of external criteria may be seen as an additional burden of the proposed procedure, attainment of such evidence is often a necessary component of the validity research agenda for high-stakes examinations.

One criticism that can be raised against the proposed procedure is that if external criteria are available, there is no need to cluster examinees. Instead, you can pick a desired value of the external criterion, calculate the average test score of examinees who attain that value, and use the median test score for this group as the passing standard (Livingston & Zieky, 1982). For

---

[3]A particular criticism of the English Composition grades used as external criteria in this study is that there were very few students with grades of "D" or "F." Thus discriminating among these students was difficult. A related problem was the proportion of examinees with missing grades. These missing data did not appear to be missing at random; more lower-achieving students did not report their grades.

example, using English Composition grades, we could select all high school seniors who were "C" students and calculate their median WST score. Although this procedure is appealing due to its simplicity, it has one major drawback: it places all of the faith in the standard setting process on the external criterion. Conversely, by clustering examinees based on their test performance, the cluster analysis procedure shares the burden of standard setting between the test instrument and the external criterion. The process begins by referencing the clustering to the test specifications (i.e., using content area subscores to cluster examinees). This step emphasizes the fact that the content areas comprising the test are meaningful, and reasserts the claim that the test, as defined by the test developers, can meaningfully separate examinees into pass/fail groupings. Identification of the ideal passing standard using cluster analysis represents a synthesis of test and criterion evaluation. An additional advantage of the cluster analysis procedure is that data from the entire sample of examinees is used to set the standard, rather than just using the data of examinees who attain a particular value on the criterion measure.

Reliability of the procedure

The stability of the clustering solution remains a primary concern in evaluating the utility of the cluster analysis for standard setting. Although the five-cluster solution exhibited stability in terms of the suggested passing standard, relationship to the external criteria, and cluster profile, some differences were also observed. Should independent panels of SMEs be selected to evaluate clustering solutions, differences in interpretation of the solutions could also emerge. Procedures to facilitate the stability of cluster solutions, and to evaluate whether adequate stability is achieved, are important areas for future research.

The present study used samples from different years to maximize the likelihood that instability would emerge. Separate analyses were also conducted after combining the 1993 and 1994 samples, and then randomly selecting two independent samples. The results of the random-group analyses provided the same conclusions gleaned from comparing the 1993 and 1994 data sets (i.e., five-cluster solution was best, the cluster profiles were similar, passing standards for the separate five-cluster solutions were identical), and so the differences noted between samples most likely reflect a "sample effect", rather than a "year effect". Given the similarities and differences noted between the separate samples, the results of the present study suggest that more than one sample be used when cluster-analyzing examinees. Looking at the results from more than one sample assists the test developers and SMEs in choosing the most appropriate passing standard, and in evaluating the stability of the solution. In any event, the effect of different samples of examinees on a clustering solution

is likely to cause less instability in the resulting passing standard than would the effect of different samples of SMEs following Angoff-type methods.

The use of cluster analysis to set standards on tests assumes that more than one homogeneous group of examinees exists in the examinee population. In some testing situations, especially in small-sample applications, this may not be an appropriate assumption. In such cases, hierarchical clustering methods may be more useful in identifying the few examinees who fall into an extreme group.

Applicability to performance-based assessments

A notable advantage in using cluster analysis to set standards on tests is that it is applicable to all types of assessments. As illustrated in the present study, the procedure is not limited to multiple-choice tests. Clustering solutions based on performance tasks are likely to separate examinees well, especially when such assessments are multidimensional. The cluster analysis procedure makes no dimensionality assumptions, although the separate scores used to cluster examinees must, of course, be reliable.

Choosing the passing standard

The choice of passing standard used in this study was based on median WST scale score for the group identified as the borderline cluster. This rationale is similar to that employed in methods using panels of SMEs to rate items. However, other strategies are possible, depending on the level of rigor desired, passing percentages, and other political or logistic factors. Some obvious options are rounding the mean borderline score up or down, or using the score +/- 1.64 standard deviation units from the mean. Other variations are also possible.

Clustering Examinees and Test Dimensionality

An interesting side observation in evaluating the results of this study was that groups that differed with respect to essay versus multiple-choice performance were observed. Previous comprehensive factor analytic studies on the WST suggested that the essay and MC sections were measuring a strictly unidimensional construct (Lukhele & Sireci, 1995). However, by focusing on examinees, rather than on test items, the differential "effects" of the essay and MC sections are directly observable. This finding may have important implications for conceptualizing test "dimensionality" and for evaluating the distinct types of information different item types provide.

Conclusion

This study represents a first step in applying cluster analysis to the standard setting problem. The results suggest that cluster analysis is an effective means for standard setting when external criteria are available. Future research should explore more compelling types of criterion variables (e.g., actual course grades rather than self-reported grades). Given the fact that external criteria are becoming requisite for test validation, this requirement is not impractical. Future research is needed to further evaluate the stability of cluster solutions and resulting passing standards. The logic of using cluster analytic procedures to set standards on tests is applicable to both multiple-choice and performance-based assessments, and is founded upon the content validity of the assessment instrument. Using cluster analysis to set standards requires interpretation and evaluation of competing cluster solutions, and of test/criterion relationships. Using expert judgment at this stage of the standard setting process may prove less subjective, and more reliable, than methods based on defining minimal competence, and rating individual test questions.

References

Anderberg, M.R. (1973). Cluster analysis for applications. New York: Academic Press.

Aldenderfer, M.S., & Blashfield, R.K. (1984). Cluster analysis. Newbury Park, CA: Sage.

GED Testing Service, (1993). GED Examiner's Manual for the Tests of General Educational Development. Washington, DC: American Council on Education.

Glass, G.V. (1978). Standards and criteria. Journal of Educational Measurement, 15, 237-261.

Green, B.F. (1981). A primer of testing. American Psychologist, 36, 1001-1011.

Livingston, S.A., & Zieky, M.J. (1982). Passing scores: A manual for setting standards of performance on educational and occupational tests. Princeton, NJ: Educational Testing Service.

Lukhele, R., & Sireci, S.G., (1995). Using IRT to combine multiple-choice and free-response sections of a test onto a common scale using a priori weights. Paper presented at the annual meeting of the National Council on Measurement in Education, April, San Francisco, CA.

Patience, W. & Swartz, R. (1987). Essay score reliability: Issues in and methods of reporting the GED Writing Skills Test scores. Paper presented at the annual meeting of the National Council on Measurement in Education, April, Washington, DC.

Shepard, L.A. (1979). Setting standards. In M.A. Bunda & J.R. Sanders (Eds.), Practices and problems in competency-based measurement (pp. 72-88). Washington, DC: National Council on Measurement in Education.

Sireci, S.G., & Biskin, B.H. (1992). A survey of national professional licensure examination programs. CLEAR Exam Review, 3, 21-25.

SPSS, (1993). SPSS for Windows, version 6.0 (computer program). Chicago, Il: Author.

Wiley, A., & Sireci, S.G. (1994). Determining the reliability of the GED Writing Skills Test using classical test theory. paper presented at the annual meeting of the Northeastern Educational Research Association, October, Ellenville, NY.

Zieky, M.J. (1994). A historical perspective on setting standards. Paper presented at the Joint Conference on Standard Setting for Large-Scale Assessments, October, Washington, DC.